LGKMA의 구조 및 특성

곽종근, 은종진, 강윤선

{kwak, zzeun, yunsun}@lgcit.com

LG종합기술원

The structure and features of the LGKMA

Jong-Geun Kwak, Zong-Zin Eun, Yun-Sun Kang

LG Corporate Institute of Technology

요약

LGKMA 시스템은 형태소 분석기와 품사 태거 및 명사 추출기로 구성되며, LG 종합기술원에서 연구 개발 중인 다국어 정보 검색, 음성 합성, 개인정보처리 에이전트 및 디지털 TV의 프로그램 안내문을 분석, 검색하는 EPG(Electronic Program Guide) 응용 등 다양한 응용 프로그램에서 사용되고 있다. 본 논문에서는 형태소 분석기와 태거의 기반 기술보다는 LGKMA(LG Korean Morphological Analyzer)의 전반적인 구조와 다른 시스템과 비교했을 때의 특성, 그리고 실제 응용되는 사례를 소개하고자 한다. 또 표준화를 위해서 열렸던 MATEC99에 참가하기 위해서 수행했던 작업들을 보고한다.

1. 서론

LGKMA 시스템은 형태소 분석기와 품사 태거 및 명사 추출기로 구성되며, LG 종합기술원에서 연구 개발 중인 다국어 정보 검색, 음성 합성, 개인정보처리 에이전트 및 디지털 TV의 프로그램 안내문을 분석, 검색하는 EPG(Electronic Program Guide) 응용 등 다양한 응용 프로그램에서 사용되고 있다. LGKMA는 다국어 정보 검색, 음성 합성 등과 같이 문장의 형태에 제한이 없는 응용 프로그램의 구성 요소로 사용되기도 하고, 디지털 TV 응용 프로그램이나 사용자의 메모 문장을 관리하는 개인정보처리 에이전트 등과 같이 입력문의 형태가 비교적 일정한 응용 프로그램의 구성요소로 사용되기도 한다.

이들 각 응용 프로그램에서 필요로 하는 LGKMA의 기능은 프로그램의 특성에 따라서 조금씩 다른데, 예를 들면 정보 검색은 높은 재현율을 얻기 위한 기능을, 음성 합성기는 자연스러운 합성음 생성에 유리한 기능을, 개인정보 처리 에이전트는 양질의 수 표현 분석 기능을 필요로 한다. 이와 같은 여러 응용 프로그램의 필요에 의해서 LGKMA에 추가된 몇 가지 특성에 관해서 살펴본다.

한편, 한국어 형태소 분석기나 한국어 품사 태거, 한국어 명사 추출기, 시소러스, 한국어 구문해석기, 영어 형태소 분석기 및 영어 품사 태거 등의 자연어 처리 핵심 요소 및 이를 이용한 정보 검색기, 개인 정보 처리 에이전트, 디지털 TV 등의 응용에 관한 개발을 계속하고 있다.

본 논문에서는 2절에서 LGKMA 시스템을 응용하는 사례에 대해서 소개하고, 3절에서 LGKMA 시스템의 구조에 대해 간략히 살펴보고 특성에 관해서 논하며, 4절에서는 형태소 분석기의 특성에 관해 살펴보고 5절에서 결론을 맺는다. 끝으로 한국어 형태소 분석기 표준화를 위해서 개최된 MATEC99에 참가하기 위해서 수행하였던 작업들과 LGKMA의 접근 방법에 관해서 소개한다.

2. LGKMA의 응용 사례

LGKMA는 정보 검색기를 비롯하여, 가계부 및 스케

쥴을 관리하는 개인 정보 처리 에이전트, 음성 합성기 등의 구성 요소로 사용되고 있다. LGKMA를 이용하는 이들 각 응용의 특성 및 이들 응용에서의 LGKMA의 역할에 대해 살펴 보기로 한다.

2.1 정보 검색기

정보 검색기는 문서를 분석하여 키워드를 추출하여 색인하고, 사용자 질의를 분석하여 색인된 문서 중에서 관련 있는 것들을 찾아 보여 준다. 그림 1에 정보 검색기의 구조를 도시하였다. 여기서 다국어 문서 분석기는 입력문이 어느 언어에 속하는지를 검사하고, 문서를 문장별로 분리하여, 해당 언어의 형태소 분석기로 넘긴다. 정보 검색기에서 LGKMA는 등록 문서 및 사용자의 질의를 분석하는 역할을 한다.

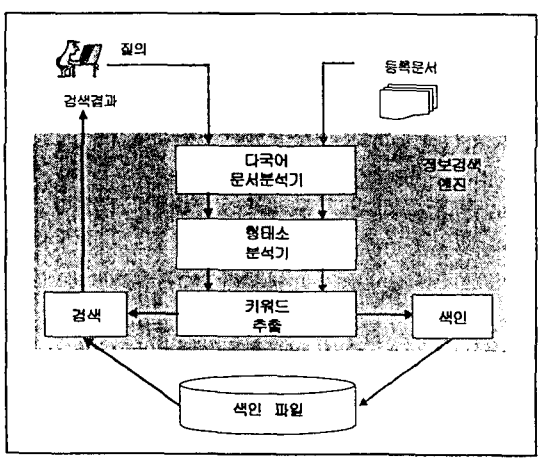


그림 1. 정보 검색기의 구조

LG에서 개발된 정보 검색기는 LG Soft의 전자 문서 관리시스템(EDMS)에 사용되고 있는데, 이 응용은 관련된 문서를 보다 많이 찾아 줄 목적으로 정확률보다는 재현율을 높이는 것을 특징으로 하고 있다. 따라서 LGKMA의 형태소 분석기에서는 많은 후보를 생성하여 옳게 분석되는 후보가 없는 경우를 줄이도록 하고 있어서 다른 형태소 분석기에 비해서 많은 분석 결과를 출력한다. MATEC99 결과에서도 명사 추출기의 재현율은 96%로 참가팀 평균 85%에 비해 매우 높은 편이었고, 정확률은 참가팀 평균 80%에 훨씬 못 미치는 63%로 나타났다.

2.2 음성 합성기

음성 합성기에서 LGKMA는 문장을 형태소 분석 및 품사 태깅을 통하여 형태론적 언어 정보를 추출하는 역할을 하는데, 이 정보는 강세(accent), 휴지(pause), 음소 지속시간(duration) 등 합성에 필요한 정보를 추출하는데 사용된다.

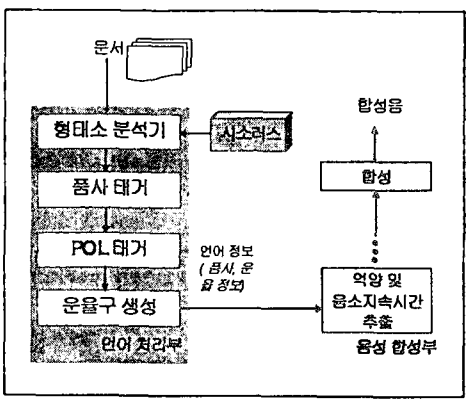


그림 2. 음성합성기의 구조

여기서 휴지나 강세 등은 문장에서 지역적(local) 정보에 의해서 결정되는 경우가 많지만, 전역적(global) 정보를 필요로 하는 경우도 있다. 따라서 보다 향상된 합성음을 얻기 위해서는 형태론적 언어 정보뿐만 아니라 전역적 정보인 문장 성분간의 수식 관계를 파악할 필요가 있다. 수식 관계를 파악하기 위해서는 구문 해석이 필요한데, 비용의 측면과 현실적인 구현의 어려움으로 인해 무제한적 도메인에서는 완전한 구문 해석기를 사용하기 힘들다는 문제가 있다. 이에 대한 해결 방안으로 명사구 중에서 인명, 조직명, 장소명 등을 추출하는 POL 태거를 이용한다. 또 POL 태거에서 전체 문장에 대한 구문 해석 과정 없이도 인명, 조직명, 장소명 등의 명사구를 정확히 추출하기 위해서, 형태소 분석기는 품사 셋을 자세히 세분하여 보다 많은 형태론적 정보를 POL 태거에 넘겨 준다.

2.3 개인 정보 처리 에이전트

LGKMA는 개인 정보 처리 에이전트의 하나인 가계부

/스케쥴 에이전트의 구성 요소로 사용되고 있다. 가계부/스케쥴 에이전트는 그림 3에 보인 것과 같이 사용자가 메모한 자연어 문장을 분석하여, 응용프로그램에 맞는 형식으로 데이터를 변환하여 응용프로그램에 저장하고, 반대로 사용자의 자연어 질의를 다시 분석하여, 응용프로그램에 저장된 데이터를 가공하여 보여주거나 수정하는 일을 한다[1].

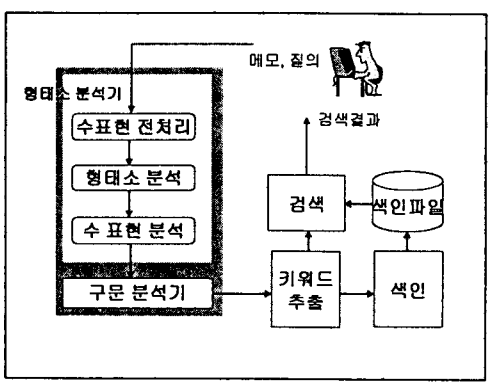


그림 3. 개인 정보 처리 에이전트

예문 (1), (2)에 보인 것처럼 가계부나 스케쥴러에서 사용되는 문장은 간단한 메모 형태로, 비록 문장의 길이는 짧지만 축약된 형태로 사용되며, 조사나 어미를 비롯하여 문장의 본동사까지 생략되는 경우가 빈번하게 발생하여 의미 해석을 어렵게 한다. 그리고 사용자가 메모한 문장과 나중에 질의한 문장이 의미는 동일하지만, 표층 요소가 달라서 올바른 검색 결과를 얻기 어려운 경우도 있다. 예문 (1)과 같이 메모 문장을 기록한 후에 (3)의 예와 같이 질의를 하는 경우, `수리비'가 `수리하는데 쓴 비용'과 의미적으로 동일하다는 것을 파악해야만 옳은 검색 결과를 낼 수 있다.

(1) 9/5 의자 수리비 만원
(2) 9/9 강남역에서 동창회
.....
(3) 질의> 의자를 수리하는데 쓴 비용은?

표 1. 가계부/스케쥴러 입력 문장의 예

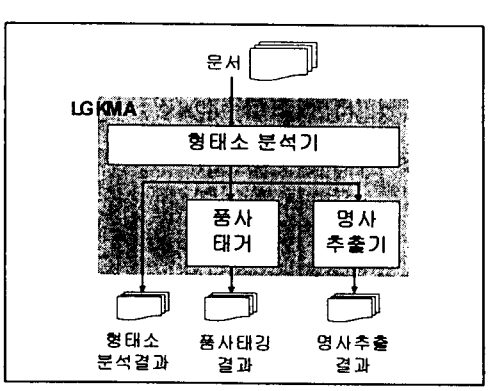
이처럼 접미사 `비'와 명사 `비용'의 의미가 동일하다는 것을 파악하기 위해서 시소러스를 사용하여, 단음절명사(NC1)나 접미사(SUFFIX) 등의 의미까지도 명확히 분석해야 하므로, 구문 분석기에서는 `수리비'의 의미를 분석하기 위해서 동작성명사(NCA)인 `수리'와 접미사 `비'의 의미 관계를 분석한다.

또한, 가계부나 스케쥴러에 사용되는 입력 문장은 수 단위와 날짜 및 금액 등의 수 표현이 빈번히 사용되는 특징이 있어서, 수 표현을 잘 처리함으로써 전체 에이전트의 성능을 향상시킬 수 있다. LGKMA의 수 표현 처리는 그림 3에서 보인 것처럼 형태소 분석의 앞과 뒤에 수표현 전처리[1]와 수표현 분석으로 구분되는데, 수표현 전처리 단계에서는 형태론적 정보를 보지 않고도 처리할 수 있는 것들을 처리하고, 수표현 분석에서는 형태소 분석된 결과를 이용하여 형태론적 정보를 보고 결정해야 할 것들을 처리한다.

이와 같이 개인 정보 처리 에이전트 응용에서는 가계부나 스케쥴 에이전트의 입력 문장 특성을 반영하여, 형태소 분석 단계에서부터의 단음절명사 및 접미사에 대한 의미 분석과, 수표현 처리를 통하여 성능을 향상시키고 있다.

3. LGKMA의 구조

LGKMA는 형태소 분석기와 품사 태거, 명사 추출기로 구성되어 있으며, 그림 4에 이를 도시하였다. 각 구성 모듈의 개괄적인 특성은 다음과 같다.



[1] 형태소 분석의 전처리 단계에서 수행하는 수표현 처리 과정을 일컬음

그림 4. LGKMA의 구조

3.1 형태소 분석기

LGKMA 형태소 분석기는 CYK 알고리즘에 기반한 차트 파싱(chart parsing) 방법을 이용하여 분석 가능한 모든 후보를 생성한다. 다른 형태소 분석기와 구별되는 특성으로는, 첫째 70여개로 확장된 품사 셋을 사용하여 형태소 분석 단계부터 가능한 한 많은 언어 정보를 추출한다는 점, 둘째 수 단위와 날짜, 시간, 금액 등(이하 수 표현)의 처리를 통해서 수 표현에서 발생하는 애매성을 줄이고, 형태소 분석 단계의 부담을 줄였다는 점, 셋째 명사 시소러스를 기반으로 한 refine 기능 등을 꼽을 수 있다. 그리고 다수의 응용프로그램의 요구를 충족시키면서 최소한의 수정으로 이용 가능하도록 범용성을 높였다는 점도 LGKMA의 특징이라 볼 수 있다.

3.2 품사 태거

품사 태거는 은닉 마코프 모델(Hidden Markov Model, HMM)을 이용한 확률적 모델을 사용한다. 확률적 모델에서는 태깅하고자 하는 문장이 주어진 언어모델에서 발화될 확률을 가장 높게 만들어 주는 품사열을 찾는 것으로, 품사가 주어졌을 때에 특정 어휘를 가질 unigram 확률값[2]과 품사의 n-gram 확률값을 연속으로 곱해서, 근사적으로 구한다[2]. 본 시스템에서 품사의 n-gram은 trigram으로 근사하여 사용한다.

LGKMA의 품사 태거를 위해서 구축한 품사 부착 학습 말뭉치는 신문, 잡지 42만여어절과 소설 1만 7천여어절로 구성되어 있다.

3.3 명사 추출기

명사 추출기는 형태소 분석 결과를 이용하여 명사를 추출한다. 재현율을 높인다는 정보 검색기의 목적에 따라서, 명사 추출기에서는 최장일치 명사 이외의 다른 분석 후보에서도 명사를 추출한다. 또 동사가 형태소 사전에 등록되지 않아서 고유명사로 분석된 후보가 명사로 잘못 추출되는 것을 막기 위해서 고유명사 중 "었", "았", "겠", "였", "습니"」 .. 등의 문자열을 포함하는 후보는 명사 추출에서 제외하는 간단한 휴리스틱을 이용하였다.

4. 형태소 분석기의 특성

LGKMA에 사용되는 형태소 분석기의 주요한 특성으로 다음과 같은 것을 들 수 있다.

3.1 확장 품사 셋 사용

LGKMA에서 사용하고 있는 품사 셋은 70여개로 세분되어 있다. 품사를 세분한다는 것은 형태소 분석기를 구현하는 측면에서나 학습 말뭉치를 유지, 보수한다는 측면에서 상당한 비용을 요구한다.

구문적, 의미적으로 완전히 구문 해석을 수행하면 덜 세분된 품사 셋을 이용하더라도 올바르게 문장을 해석할 기회가 있지만, 응용 프로그램에 따라서 구문 해석기를 사용하는 것이 비효율적인 경우가 많다. 또한 제한되지 않은 일반적인 도메인에서 사용 가능한 구문해석기는 현실적인 구현의 어려움으로 실용화해서 사용하는 사례가 드물다. 그 대신에 형태소 분석기나 품사 태거, 혹은 간단한 shallow parser 등을 이용하여 시간이나 날짜 표현 등과 같은 명사류를 처리하는 응용들이 대부분이다. Shallow parser에서는 구문적, 의미적으로 완전히 분석하는 것이 아니므로, 형태소의 품사 정보가 세분화되어 더 많은 정보를 이용할 수 있으면 구현이 쉬워지고 성능향상을 꾀할 수 있다.

품사 태거의 경우에도 확장된 품사 셋을 사용함으로써 n-gram 품사 확률을 계산할 때에 더 많은 정보를 이용하여 계산할 수 있다. 예를 들어 LGKMA 시스템에서는 조사의 품사 셋을 부사격조사(JCA), 접속격조사(JCJ), 병렬접속격조사(JCJP), 관형격조사(JCM), 목적격조사(JCO), 인용격조사(JCQ), 주격조사(JCS), 호격조사(JCV), 보조사(JX)의 9가지로 세분해서 사용한다. 이와 같이 세분하여 사용함으로써 '물이 차다' 와 '공을 차다' 라는 예문

[2] P(w|t), w는 어휘, t는 품사를 나타냄

에 대해서 `차다'의 품사를 선택할 때에 주격조사와 목적격조사에 대해 서로 다른 품사의 n-gram 확률을 사용하여 변별력을 높일 수 있다.

(3) 물이 차다 → 물/NC1+이/JCS 차/PA+다/EF
→ JCS를 포함하는 n-gram 확률이 사용됨
(4) 공을 차다 → 공/NC1+을/JCO 차/PV+다/EF
→ JCO를 포함하는 n-gram 확률이 사용됨
(NC1:단음절명사, PV:동사, PA:형용사, EF:종결어미)

표 2. 품사 태거 입력 문장의 예

이처럼 품사 셋을 확장하여 품사 태거에 적용하면, 분류가 자세할수록 data sparseness 문제해결을 위해서 학습 말뭉치의 크기를 늘려주어야 하지만, 학습 말뭉치에서 학습할 때에나 문장의 확률을 계산할 때에 더 많은 클래스를 이용하여 분류할 수 있음을 의미한다.

3.2 수 표현 처리

수 표현 처리의 목적은 한국어 문장을 분석할 때 수 표현에 대한 형태소 분석 과정에서 발생하는 애매성을 효과적으로 제거하고, 수 표현의 의미를 정확히 분석하는 데 있다.

LGKMA의 수 표현 처리기에서 사용하는 수단위 사전에는 단위성 의존명사가 어떤 종류의 수사와 결합 가능한지에 대한 정보, 단위성 의존명사와 함께 올 수 있는 명사의 의미적 분류, 단위성 의존명사로 형성된 수 단위와 명사와의 관계 및 수 단위와 명사가 결합한 결과로 새로 생기는 의미에 관한 정보가 저장되어 있다.

수 표현 처리 과정은 그림 3에 보인 것과 같이 형태소 분석의 전처리 부분과 형태소 분석한 결과에 대한 수 표현 분석 과정으로 나누어 진다. 수표현이 처리되는 유형을 살펴 보면 크게 세가지로 구분할 수 있다. 첫째, 예문 (5)와 (6)과 같이 애매성이 없어서 수 표현 처리[3]만 하고 형태소 분석은 하지 않는 것과, 둘째, 예문 (7), (8)과 같이 수표현 전처리기에서는 애매하지만, 형태소 분석을 한 후에 수표현 분석 단계에서는 애매성을 제거할 수 있는 것과, 셋째, 예문 (9)와 같이 수 표현 전처리기 및 형태소 분석, 그리고 수 표현 분석 과정을 거쳐도 여전히 애매성이 남아 있는 것이 있다. 표 3에 수 표현 전처리를 수행한 결과를 보였다.

(5) "1:20" → 01:20/TIME
(6) "99/7/20~99/8/30"
→ y-1999 m-7 d-20/DATE + ~/SYMBOL_TILDE
+ y-1999 m-8 d-30/DATE
(7) "작년 봄에"
→ (작년 봄, y-1998 s-SPRING/DATE)에
(8) "작년 봄옷"
→ (작년 봄, y-1998 s-SPRING/DATE)옷
(9) "수영장 만원"
→ 수영장 (10000 원/UNITN, 만원)
이 경우, 형태소 분석 단계에서는 "만원"에 대한 형태소 분석 결과와 전처리로부터 넘겨 받은 수 단위 토큰 `10000 원/UNITN'을 분석 결과로 출력한다. 여기서 UNITN은 숫자 수단위를 의미한다.

표 3. 수 표현의 전처리 결과 예

수 표현 분석 단계에서는 예문 (6)과 같이 전처리된 다중 토큰이 다시 하나의 수 단위로 묶일 수 있는 것을 묶어 주고, 예문 (7), (8) 등과 같이 애매성을 제거할 수 있는 것들에 대해서 불필요한 후보를 삭제한다. 즉 예문 (8)과 같이 수 단위 토큰 뒤의 형태소가 `봄옷'처럼 한 어절 복합명사를 이루는 것은 수 단위 처리 결과를 삭제한다.

이과 같은 수 표현 처리를 통하여 형태소 분석 단계의 부담을 덜어 주고, 다중 어절의 많은 정보를 보고 처리한 수 표현에 대해서는 분석 결과의 애매성을 줄이는 효과도 기대할 수 있다. 특히 가계부/스케쥴 에이전트 등과 같은 응용에서는 입력문에 수 표현이 빈번하게 발생하여, 수 표현의 적절한 처리가 에이전트의 성능을 향상시킨다.

[3] 수표현 전처리 및 수표현 분석

3.3 시소러스 개념을 이용한 refine

LGKMA 형태소 분석기는 형태소 분석 결과를 시소러스의 의미 관계를 사용하여 refine 한다. 이러한 refine 과정은 의미적 결합 정보를 시소러스의 개념을 이용하여 규칙으로 표현하고, 분석된 결과가 이러한 규칙에 맞는지를 시소러스 개념 상의 통합(unification)연산[4]을 통하여 검증함으로써 이루어진다.

표 4는 refine을 위한 정보 중 일부를 나열한 것이다. LGKMA에서 사용하는 refine 정보는 의미 기준으로 약 180여개가 있고, 이 중 40여개는 단음절 명사에 의한 정보이고 140여개는 접미사에 관한 정보이다. 이 정보를 바탕으로 다음의 긍정적 규칙(Positive rule)과 같은 부정적 규칙(Negative rule)을 이용한다.

1. Positive Rule : Refine 정보에 언급된 단음절명사, 접미사의 앞에 오는 형태소의 품사가 refine 정보에 접속 가능 품사로 표시되어 있거나, 앞에 오는 형태소의 의미가 접속 가능 의미에 해당하면 이 후보는 우선시 된다.
2. Negative Rule : 고유명사(NQ), 인명을 나타내는 품사(NLNAME, NFNAME), 외래어(F) 등과 같이 단음절명사, 접미사와 결합하여 애매성을 가중시키는 특정 품사는 refine 정보에 의해 접속이 허락된 일부 단음절명사, 접미사와의 결합만 허용하고 나머지는 삭제한다. 예를 들어 표 1의 refine 정보에서 `각 a'는 고유명사와는 접속을 허락함을 보이고 있다. `간 a', `간 5', `간 b' 등은 고유명사와 접속할 수 없다.

[4] 시소러스 관계 R에 대한 시소러스 개념 상의 통합 연산은, 관계 R에 대해서 개념 C1이 개념 C2의 하위 개념일 때에 R에 대해 C1은 C2에 통합된다고 정의한다. LGKMA의 시소러스에서 사용하는 관계는 개념의 상,하위어 관계(Hypernym/Hyponym), 구성 요소의 상하위어 관계(Component Holonym/Meronym), 멤버의 상하위어 관계(Member Holonym/Meronym), 장소의 상하위어 관계(Place Holonym/Meronym)로 세분되며, 이 4가지 관계 중에서 개념 C1이 개념 C2에 통합되는 것이 하나라도 존재하면 C1은 C2에 통합되는 것으로 정의한다.

단음절/접미사	접속 가능 품사	접속 가능 의미
..	..	..
각 a(누각)	NQ	Don't care
간 a(기간)	DATE, TIME	기간 1
간 5(관계)	Don't care	관계 1
감 b(사람)	Don't care	사람 1
..	..	..

DATE, TIME: 전처리에 의해서 날짜와 시간으로 묶인 것

표 4. Refine 정보(일부)

표 5는 `부부간' 과 `한달간' 에 대한 형태소 분석 결과에서 각각 접미사와 단음절명사의 후보가 refine 과정을 통해 삭제된 후의 결과를 나타내고 있다.

```
Input:부부간
부부/NC(부부 2)+간/NC1(간 5)
Input:한달간
한달/DATE(달 2)+간/SUFFIX(간 a)
```

표 5. Refine되는 형태소 분석 예문

여기서 시소러스에 등록된 간의 의미는 간 a, 간 b, 간 2 간 4, 간 5 의 5개이다. 여기서 간 a, 간 b 처럼 뒤에 영문으로 태그가 붙은 의미는 형태론적으로 접미사임을, 간 2 간 4, 간 5 등과 같이 뒤에 숫자가 붙은 것은 (단음절)명사임을 나타낸다. 표 6은 시소러스의 내용 일부 중 관련된 항목을 보이고 있다.

```
간 a < 기간 1 < 때 1 < ..
간 5 = 사이 4 < 관계 1 < 추상 1
부부 2 < 관계 1 < 추상 1
  = 는 동의어 관계를 나타내고, < 는 상하위어 관계를 나타낸다. 편의상 관계이름은 생략한다.
```

표 6. 시소러스 내용(일부)

시소러스 개념 상의 통합 연산은 비용이 많이 들고, 이를 이용하여 refine된 결과를 출력하기 위해서 형태소 분석기는 형태론적 정보에 의미 정보까지 관리하는 부담을 안게 된다. 그러나 의미적으로 refine 함으로써 형

태론적 정보만을 이용할 때에 형태소 분석 결과가 과다 생성(overgeneration)되는 것을 막을 수 있다는 장점과, 형태소 분석기의 출력에 한번 refine 된 의미 정보를 포함하고 있으므로 구문 해석의 과정 없이 형태소 분석 자체의 결과만으로도 의미적 해석이 어느 정도 가능하다는 장점이 있다. 또 구문 해석을 이용할 경우에도 애매성이 훨씬 줄어든 상태에서 해석을 시작할 수 있다.

LGKMA 에서는 refine 을 위한 부담을 최소한으로 줄이면서 효과를 높이기 위해서 형태론적으로 애매성이 높은 단음절 명사와 접미사를 중심으로 규칙을 만들고 이들에 대해서 refine 을 수행한다. 또, 시소러스 개념 상의 통합 연산을 효율적으로 수행하기 위해서 개념들을 연산에 편리한 코드 형태로 미리 컴파일하여 사용함으로써 통합 연산에 대한 부담을 줄이고 있다.

5. 결론

LGKMA 시스템은 확장된 품사 셋을 사용하여 상위 응용 프로그램들에게 더 많은 정보를 제공하며, 수 표현에 대한 처리를 통하여 한국어 입력문에서 빈번하게 나타나는 수사 처리의 애매성을 최대한 줄였다. 그리고 형태소 분석 단계에 시소러스의 의미를 이용한 refine 을 수행함으로써 부적합한 형태소 분석 결과를 상당 부분 걸러낼 수 있었다.

6. MATEC99 참가를 위해 했던 작업

MATEC99 에 참가하기 위해서는 서로 다른 품사 셋에 대한 사상(mapping)문제를 해결해야 했다. 여기에 대한 방법으로는 기존 엔진을 ETRI 품사 셋에 맞게 바꾸는 방법과, 기존 엔진은 그대로 유지하면서 품사 셋만 사상(mapping)하는 두 가지 방법이 있다.

LGKMA 에서의 접근 방법은 기존의 엔진은 가능한 한 수정하지 않고 ETRI 측의 결과를 낼 수 있도록 하는 것이었다. 따라서 LGKMA 의 분석 결과를 ETRI 품사 셋으로 사상하는 과정이 필요했고, 학습용 말뭉치에 대해서는 다음과 같은 세가지 방법을 생각해 볼 수 있었다.

첫째, 기존에 사용하던 학습용 말뭉치만 사용하고 ETRI 측에서 제공한 학습 말뭉치를 전혀 사용하지 않는 방법. 태깅된 결과만 ETRI 품사셋으로 사상한다. 둘째, ETRI 품사 셋을 그대로 학습시키고, 형태소 분석 결과를 먼저 ETRI 셋으로 사상한 후 태거를 수행하는 방법. 세째, 말뭉치를 LGKMA 의 품사 셋으로 사상하고[5] 이를 이용하여 형태소 분석기, 품사 태거를 수행한 후에 ETRI 품사 셋으로 사상하는 방법이 있다.

첫번째 방법은 가장 손쉬운 방법이긴 하지만, 대회에 참가해서 얻을 수 있는 잇점이 없다는 점에서 고려대상에서 제외되었고, 두번째 방법에 비해 세번째 방법은 가장 많은 노력을 필요로 하는 방법이지만, LGKMA 에 맞는 학습된 말뭉치를 비교적 쉽게 얻을 수 있고, 세분된 품사 셋에 의한 학습과 확률 계산이 태거의 품사 선택에 유리하다는 점에서 이 방법을 따르게 되었다. 이에 해당하는 예는 앞절의 "공을 차다"와 "물이 차다"에서 볼 수 있다.

그리고 이 방법에 따라 LGKMA 의 품사 셋으로 변환된 학습 말뭉치에 존재하면서, 우리 사전에 등록되지 않은 단어들 중의 일부를 사전 엔트리로 추가했다.

LGKMA 의 품사 셋이 superset 이므로 LGKMA 의 품사 셋을 ETRI 의 품사 셋으로 바꾸는 사상 문제는 간단했다. 그러나 어휘에 따라서 사전에 등록 여부가 다르고, 분석되는 모양이 다른 것들이 존재했다. 이와 같은 것들의 처리는 정규화된 규칙이 존재하지 않으므로 case by case 의 변환 규칙 데이터베이스를 만들고, 이를 검색하는 수준에서 해결할 수 밖에 없었고, 이 규칙들도 빈번히 발생하는 순서로 등록하였기 때문에 매우 불완전하였으며, 일관성이 결여되었다는 점이 아쉬운 점으로 남는다. 다음 대회 때에는 이 같은 점에 대한 보완 및 대책이 있기를 기대한다.

---

[5] LGKMA 의 품사셋이 superset 이므로 주위 문맥에 따라서 사상을 자동화 할 수 있는 부분은 자동화하고, 자동화가 안되는 부분만 사람이 수작업으로 사상했다.

참고서적

[1] 견종서, 서병락, 은종진, 강윤선, 백은옥, "메모의 분석과 정보 추출을 이용한 일정관리", HCI '97 학술대회 발표 논문집, 273-278, 1997

[2] E. Charniak, "Statistical Language Learning", The MIT Press, 1993